# KEIGO-SYNC : Evaluating the Impact of Politeness Levels on Factual Reliability in Japanese Large Language Models

**Aubourg Thomas, Demeude Edgar, Dupuy Guilhem and Vu Anh Duy**

## 1 Introduction

Sycophancy in large language models (LLMs)—the tendency to agree with user-stated beliefs conflicting with verifiable facts—constitutes a persistent alignment challenge [8]. While existing research examines sycophantic behavior primarily in English [1], the role of grammatical register as a systematic factor remains insufficiently explored, particularly in high-context languages where politeness is grammatically encoded [5].

KEIGO-SYNC introduces a controlled evaluation framework measuring the relationship between grammatical politeness and sycophantic agreement using Japanese, where honorific speech (Keigo) encodes social deference through obligatory syntactic structures. Each prompt is instantiated across four register conditions: English Neutral, Japanese Tameguchi (casual), Teineigo (polite), and Sonkeigo (honorific), enabling empirical assessment of whether linguistic deference correlates with measurable changes in model agreement behavior.

## 2 Dataset Construction

The corpus comprises 60 prompts designed to isolate the effect of grammatical politeness on factual integrity. To ensure diversity and comparability, the dataset integrates items from the Logical Fallacy Detection[4] benchmark, the ELEPHANT dataset, and original questions specifically developed for this study. These items are organized into five thematic domains: education, health, logical fallacies, technology, and ethics, drawing on thematic categories established in benchmarks like MMLU [3] and the Logical Fallacy Detection dataset [4] Topics are restricted to widely documented general knowledge to limit specialized knowledge biases.

All items follow a unified [Assertion] + [Question] template, formulated assertively and deliberately incorporating erroneous premises or reasoning biases to measure models' ability to maintain critical stance under linguistic pressure. Originally written in English, all prompts were adapted by a native Japanese speaker to ensure semantic and pragmatic fidelity. Each item was instantiated across three Japanese registers: Tameguchi (informal), Teineigo (standard), and Sonkeigo (honorific). By treating politeness level as the sole independent variable, this design enables controlled analysis of how linguistic deference influences the rate of sycophantic agreement versus factual correction.

## 3 Experimental Setup

### *3.1 Model Selection and Configuration*

We evaluated models of varying origins and scales: Asian-origin (Qwen 2.5 7B, GLM-4.6V-Flash) versus Western-origin (Gemma 3 27B, Mistral v0.3 7B), spanning 7B to 27B parameters. Mistral 7B base served as a non-RLHF baseline. All inferences used temperature 0.5 with N=3 iterations.

### *3.2 LLM-as-Judge Framework*

Factual correctness and sycophantic behavior are evaluated using an LLM-as-judge framework with dedicated bilingual English–Japanese judging models. The primary evaluation is conducted using Qwen 2.5-7B Instruct and Meta's Llama 3.2 3B-Instruct. To ensure validity and alignment, these judgments are complemented by Google's Gemma 3 (27B) on a 15% subsample and human verification on a 10% subsample. All judges are fully isolated from the answer-generation models and have no access to the trap construction process, prompt variants, or model identities, ensuring evaluations rely solely on the observable content of the responses.

Each evaluation instance provides the judge with an English factual reference answer, the model response, and structured scoring instructions. The English reference anchors the evaluation to a consistent factual baseline and reduces cross-lingual ambiguity. Judges assess semantic content and factual alignment, ignoring surface-level stylistic or politeness differences unless they interfere with factual accuracy.

### *3.3 Evaluation Metrics*

#### *3.3.1 LLM-as-Judge*

Sycophancy is assessed using a structured evaluation prompt operationalizing three dimensions derived from the ELEPHANT framework[2]: **regressive sycophancy** (sacrificing factual accuracy to conform to user's incorrect beliefs), **validation sycophancy** (emphasizing emotional reassurance over objective guidance), and **framing sycophancy** (uncritically accepting user's assumptions rather than correcting them). Each dimension is scored 1–5, with higher values indicating greater user appeasement over truthfulness. The final sycophancy score is calculated as the arithmetic mean of these three dimensions. Judges compare model responses against English

reference answers treated as factual ground truth, distinguishing between responses that correct misleading statements, partially align with truth, or fully adopt erroneous premises. Brief politeness is excluded from high sycophancy ratings, whereas repeated validation or uncritical adoption of misleading premises is penalized. A separate Qwen Instruct instance parses judge outputs, extracting the three dimension scores, computed average, and textual rationale into standardized JSON format, ensuring reproducible and interpretable evaluations.

#### 3.3.2 *Quantitative Metrics*

To quantify the trade-off between social deference and factual accuracy, the system employs a custom metric engine computing two primary variables. The **Formality Ratio** measures response politeness (normalized 0.0–1.0) via rule-based detection of Japanese honorific markers: Sonkeigo verbs (e.g., *meshiagaru*, *zonjimasu*, 1.0), Teineigo copulas (*desu/masu*, 0.6), and Tameguchi particles (*daro*, *jan*, 0.2). The **Factual Adherence** quantifies semantic alignment by embedding the Japanese response using Gemini Embedding 001 and calculating cosine similarity against the English ground truth embedding. By correlating Formality Ratio with Factual Adherence, the framework identifies sycophantic drift—cases where models sacrifice factual accuracy to comply with high-context politeness constraints—enabling precise quantitative assessment of how linguistic deference influences truthfulness.

## 4 Evaluation Protocol

### 4.1 *Human Evaluation of LLMs as a Judge*

To validate the efficacy of the LLM-as-a-judge framework, the automated metrics were benchmarked against human judgment, following established protocols [7]. This validation utilized a representative subsample of the dataset ($n = 72$ responses, derived from 18 of the 60 primary questions across four distinct models), accounting for 7.5% of the total response corpus. To mitigate potential systemic bias, the human evaluation was conducted by native Japanese speakers external to the research project, ensuring an objective baseline.
The protocol prioritizes the alignment of evaluative trends over absolute score magnitude. Specifically, the framework assesses the inter-rater reliability between human judges and individual LLM judges through two statistical lenses: Spearman's rank correlation coefficient, to evaluate the consistency of model rankings by quality, and Pearson's correlation coefficient, to quantify the linear relationship and magnitude alignment between the scores.

### 4.2 *Statistical Analysis and Robustness Checks*

The following table presents the statistical alignment between human evaluators and the three LLM judges regarding the *framing* metric.

**Table 1:** Inter-rater Reliability: Spearman and Pearson Correlations for Framing

| | Judge model | | |
|---|---|---|---|
| Metric | Gemma 3-27B | Llama 3.2-3B | Qwen 2.5-7B |
| **Pearson Linear ($r$)** | | | |
| Coeff. | **0.4539** | 0.3413 | 0.1790 |
| $p$-value | (0.0006) | (0.0115) | (0.1952) |
| **Spearman Rank ($\rho$)** | | | |
| Coeff. | **0.4075** | 0.2986 | 0.1650 |
| $p$-value | (0.0022) | (0.0283) | (0.2332) |

The empirical data indicates a lack of statistically significant correlation for the **Qwen 2.5-7B Instruct** model ($p > 0.19$ for both coefficients). Consequently, this model was excluded from subsequent analysis as its evaluative output does not reliably reflect human judgment. In contrast, both **Gemma 3-27B** and **Llama 3.2-3B Instruct** demonstrate a positive correlation with human ground truth. While the correlation coefficients remain moderate ($r \approx 0.30$–$0.45$), the remarkably low $p$-values ($p < 0.05$) confirm that these associations are statistically robust and not the result of stochastic variation.
The degree of alignment with human evaluators also appears to scale positively with model parameters and general reasoning capabilities. Furthermore, a significant divergence in absolute score magnitude was observed. Human evaluators assigned a mean sycophancy score of 0.80, whereas **Gemma 3-27B** yielded a mean of 2.09. This discrepancy suggests a fundamental difference in evaluative calibration: while the trends remain correlated, LLMs exhibit a higher sensitivity to sycophantic markers, resulting in more "stringent" or "harsh" grading behaviors compared to human subjects.

## 5 Results

The empirical evaluation reveals a consistent relationship between grammatical politeness and sycophantic behavior, governed by both linguistic framing and model architecture. Table 2 presents the aggregate sycophancy scores across language variants, while Table 3 details the performance of individual generation models.

### 5.1 *Impact of Language and Register*

The transition from English to Japanese emerged as the strongest predictor of sycophantic behavior. The mean absolute difference between the English baseline and Japanese variants ($\Delta_{\text{EN}\to\text{JP}} = 0.47$ for Gemma) was over twice the magnitude of internal register variations, indicating that the language shift itself drives the bulk of the behavioral change.
**The "Sonkeigo" Effect.** Crucially, the hyper-formal **Sonkeigo** register consistently yielded the highest sycophancy scores across all evaluation conditions, validating the hypothesized politeness-truthfulness trade-off. Gemma 3-27B recorded a 42.4% increase from the English baseline to Sonkeigo, while Llama 3.2-3B showed a 21.9% increase. Most notably, human evaluators demonstrated a 45.2% spike, confirming that excessive grammatical deference correlates strongly with increased agreement.
**Non-Linearity and Calibration.** While the extremes (English vs. Sonkeigo) followed a predictable pattern, the intermediate **Teineigo** register defied monotonic expectations. Both Human and Gemma 3 evaluators rated standard politeness (Teineigo) lower than casual speech (Tameguchi), with human scores dropping from 0.88 to 0.79. Regarding calibration, although human evaluators applied stricter standards (mean = 0.80) than LLM judges (Gemma: 1.75, Llama: 2.12), all judges preserved the fundamental ordering of English $<$ Japanese, confirming the robustness of these findings.

**Table 2:** Average sycophancy scores by judge and language variant

| | Judge model | | |
|---|---|---|---|
| Language variant | Human | Gemma 3-27B | Llama 3.2-3B Instruct |
| EN_Base | 0.62 | 1.39 | 1.87 |
| JP_Tameguchi | 0.88 | 1.82 | 2.11 |
| JP_Teineigo | 0.79 | 1.79 | 2.23 |
| JP_Sonkeigo | **0.90** | **1.98** | **2.28** |

### 5.2 Architectural Susceptibility

Disaggregating results by generation model reveals significant disparities in robustness (Table 3). **Mistral v0.3** demonstrated the highest overall susceptibility, recording an extreme peak of **2.57** in the Sonkeigo register: the highest individual score observed in our study. It also displayed the most pronounced "language gap," with scores surging by +1.04 points (1.53 $\rightarrow$ 2.57) upon shifting to Japanese. In contrast, **Gemma 3-27B** exhibited the greatest relative robustness, maintaining consistent scores (max 1.88) across registers, suggesting that its larger parameter count or specific alignment training provides better resistance to honorific-induced bias.

**Table 3:** Average sycophancy scores by generation model across language variants (aggregated across all judges)

| | Generation Model | | | |
|---|---|---|---|---|
| Variant | GLM-4 | Gemma 3 | Qwen 2.5 | Mistral |
| EN_Base | 1.62 | 1.55 | 1.45 | 1.53 |
| JP_Tameguchi | **1.99** | **1.88** | 1.93 | 2.29 |
| JP_Teineigo | 1.95 | 1.86 | 1.95 | 2.28 |
| JP_Sonkeigo | 1.96 | **1.88** | **2.01** | **2.57** |

**Training Methodology and Register Sensitivity.** These results highlight that sycophancy is an interaction between linguistic cues and model-specific priors. The extreme sensitivity of **Mistral v0.3** can be attributed to its training methodology: unlike Gemma 3 or Qwen 2.5, which benefit from extensive RLHF and multilingual optimization, Mistral lacks aggressive safety alignment in non-primary languages. Consequently, it likely associates high-honorific tokens with a "service persona," conflating politeness with total submission.

Conversely, **GLM-4** deviated from the standard trajectory, peaking in the casual *Tameguchi* register (1.99) rather than Sonkeigo. This inverted pattern suggests that distinct training distributions can encode sycophancy in unexpected registers, independent of standard politeness hierarchies.

## 6 Conclusion

This study introduces KEIGO-SYNC, a controlled evaluation framework for measuring the relationship between grammatical politeness and sycophantic behavior in multilingual large language models. By isolating honorific register as an independent variable while maintaining semantic invariance, we provide empirical evidence for a language-level sycophancy gap that substantially exceeds intra-register variation within Japanese.

### 6.1 Principal Findings

Our results demonstrate two primary phenomena. First, a consistent cross-linguistic effect emerges whereby Japanese prompts elicit systematically higher sycophancy scores than semantically equivalent English prompts, with language transition effects ($\Delta_{\text{EN}\rightarrow\text{JP}}$) approximately 2–3 times larger than maximum intra-Japanese register differences. Second, while Sonkeigo (honorific) register exhibited the highest sycophancy scores in all of our judges conditions, the gradation across Japanese registers (Tameguchi, Teineigo, Sonkeigo) displayed non-monotonic patterns suggesting complex interactions between grammatical formality, cultural priors embedded in training data, and model-specific alignment objectives. This aligns with recent findings that Japanese linguistic formalities can significantly alter the quality and accuracy of AI-generated technical content [6].

### 6.2 Limitations and Methodological Constraints

Several methodological limitations constrain the generalizability of these findings and warrant explicit acknowledgment.

**Model Constraints.** The LLM-as-judge evaluation relied on models of limited capacity (Gemma 3-27B, Llama 3.2-3B) due to budgetary and computational constraints. Ideally, frontier models such as Gemini 3 or GPT-5 would serve as judges to maximize linguistic sophistication and cross-cultural reasoning capability. The use of smaller open-source models, while enabling reproducibility, may have introduced evaluation noise or failed to capture subtle semantic nuances in Japanese honorific constructions.

**Dataset Scale.** The evaluation corpus comprised only 60 prompts, falling substantially short of the initially projected target of 200+ items. This limited sample size constrains statistical power for detecting interaction effects between register, semantic domain, and model architecture. Fine-grained analyses of domain-specific or fallacy-type-specific sycophancy patterns remain underpowered, and confidence intervals around effect size estimates are correspondingly wide.

**Prompt Design Uniformity.** All experimental items followed a rigid [Assertion] + [Question] template structure. This uniformity, while controlling for structural confounds, limits ecological validity. Real-world sycophancy traps exhibit substantial variation in rhetorical framing, interrogative vs. declarative phrasing, and implicit vs. explicit false premises. The absence of paraphrastic variation and alternative formulation strategies may have artificially stabilized model responses, potentially underestimating the true variance in sycophantic behavior across naturalistic interaction contexts.

**Translation Quality Assurance.** While all Japanese translations were produced by a native speaker, the absence of formal linguistic expertise in Japanese honorific pragmatics introduces potential validity threats. Subtle errors in Keigo application, inappropriate register mixing, or unnatural formality gradations may have inadvertently confounded the manipulation. Ideally, translations would undergo expert review by a specialist in Japanese sociolinguistics or undergo inter-rater reliability assessment by multiple native speakers with metalinguistic training.

### 6.3 Future Directions

Despite these limitations, KEIGO-SYNC establishes a replicable methodology for isolating grammatical register effects in safety-critical model behaviors. Future work should expand the dataset to include additional high-context languages (e.g., Korean, Thai, Javanese), increase sample sizes to enable robust statistical inference, diversify prompt structures to capture naturalistic variation, and employ frontier judge models or specialized bilingual annotators. Additionally, controlled ablation studies examining the interaction between RLHF intensity and honorific-induced sycophancy could clarify whether this phenomenon is intrinsic to alignment procedures or correctable through modified training objectives. Ultimately, addressing the politeness-truthfulness trade-off represents a necessary condition for the safe deployment of culturally-aware language models in global settings.

## 7 Citations and references

## References

[1] J. Armengol-Estapé, F. Ladhak, C. Bonet, et al. Quantifying multilingual performance of llms across languages. *arXiv preprint arXiv:2404.11553*, 2024.

[2] M. Cheng, S. Yu, C. Lee, P. Khadpe, L. Ibrahim, and D. Jurafsky. Elephant: Measuring and understanding social sycophancy in llms. *arXiv preprint arXiv:2505.13995*, 2025.
[3] D. Hendrycks, C. Burns, S. Basart, et al. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
[4] Z. Jin, A. Lalwani, T. Vaidhya, X. Shen, Y. Ding, Z. Lyu, M. Sachan, R. Mihalcea, and B. Schölkopf. Logical fallacy detection. *arXiv preprint arXiv:2202.13758*, 2022.
[5] P. Liang, R. Bommasani, T. Lee, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.
[6] K. Matsui et al. The honorific effect: Exploring the impact of japanese linguistic formalities on ai-generated physics explanations. *arXiv preprint arXiv:2407.13787*, 2024.
[7] E. Perez, S. Ringer, K. Lukošiūtė, K. Nguyen, E. Chen, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.
[8] M. Sharma, M. Ashton, M. Glukhov, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.